Andrew Young School of Policy Studies
Research Paper Series

Working Paper 07-26
May 2007

Department of Economics

Patents, R&D and lag effects: Evidence from flexible methods for count panel data on manufacturing Firms

Shiferaw Gurmu
Georgia State University

Fidel Perez-Sebastian
Univesidad de Alicante

This paper can be downloaded at:
http://aysps.gsu.edu/publications/2007/index.htm

The Social Science Research Network Electronic Paper Collection:
http://ssrn.com/abstract=990025

Georgia State University | Andrew Young School of Policy Studies

Patents, R&D and lag effects: Evidence from flexible methods for count panel data on manufacturing Firms*

Shiferaw Gurmu†
Department of Economics
Andrew Young School of Policy Studies
P.O. Box 3992
Georgia State University
Atlanta, GA 30302

Fidel Perez-Sebastian
Schoool of Economics
Univesidad de Alicante
03690 Alicante, Spain

This Version: March 2007

Abstract

This paper investigates the relationship between patents and research and development expenditures using new longitudinal patent data at the firm level for the U.S. manufacturing sector from 1982-1992. The paper also develops a new class of count panel data models based on series expansion of the distribution of individual effects. Estimation results from various distributed lag and dynamic multiplicative panel count data models show that the contemporaneous relationship between patenting and R&D expenditures continues to be strong, accounting for over 60% of the total R&D elasticity. The lag effects are higher than have previously been found for the 1970's data.

Keywords: Innovative activity. Patents and R&D. Individual effects. Count panel data methods.
Field Designation: C20. O30.

*We would like to thank Chris Bollinger and Paula Stephan for useful comments on the previous version of the paper. Fidel Perez-Sebastian acknowledges financial support from the Spanish Ministry of Science and Technology, SEJ2004-08011ECON, and by the Instituto Valenciano de Investigaciones Económicas. Earlier versions were presented at the 11*th* International Conference on Panel Data, Texas A&M University, the Midwest Econometrics Group Meeting, and the Annual Conference of the Southern Economic Association.

†Correspondence to: Shiferaw Gurmu, E-mail: sgurmu@gsu.edu, phone: (404) 651-1907, and fax: (404) 651-4985.

1 Introduction

This paper investigates the relationship between patents and research and development (R&D) expenditures using new longitudinal patent data at the firm level for the U.S. manufacturing sector from 1982-1992. We estimate various distributed lag and dynamic multiplicative panel count data models, and compare results from the new patent data set to results from the sample first analyzed by Hall, Griliches and Hausman (1986) that cover the 1970's (henceforth, called HGH data). The paper also develops a new class of count panel data models based on series expansion of the distribution of unobserved heterogeneity. The model proposed may be thought of as a semiparametric generalization of the negative binomial and beta mixture model of Hausman *et al.* (1984).

The patents and R&D relationship has attracted enormous attention in the literature. The reason is a powerful one: innovative activity at the firm level is important for firms to improve their performance, and is the main driving force of the growth process in advanced economies. In the cross-section and time-series dimension, the basic approach is to estimate a knowledge production function that converts current and past R&D investment into patents that are taken as an output measure of the inventive process.[1] The main goal is to infer from the lag distribution on past R&D something about gestation lags in knowledge production.

Estimating this knowledge production function is, however, no easy task. Starting with the seminal work by Hausman *et al.* (1984), several count data models have been proposed

[1] See Griliches (1990) for summary of the literature on the use of patents as economic indicators to understand the process of innovation and technical change. One of the problems in using patents as an outcome variable is that not all innovations are patented and patents differ in their economic impacts. However, there is some evidence that patents provide a fairly reliable, although not perfect, measure of innovative activity at the industry level (Acs and Audretsch 1989; Griliches 1990). See Stephan *et al.* (2000) for discussion of unit of analysis and the spillover process.

and employed extensively to analyze the relationship between patents and R&D expenditure employing both cross section and panel data.[2] The special features that patent data exhibit is what makes it quite a challenge to develop statistically satisfactory and economically meaningful models; see, for example, Guo and Trivedi (2002). Patent counts display heavy upper tails, relatively low median values, relatively high means, and substantial proportion of zero patents typically coinciding with the mode of the patent distribution. In particular, the large degree of skewness in patent distribution may be attributed to the presence of observed factors (such as R&D expenditures and firm size), unobserved heterogeneity (such as differences in quality of patent innovations), and other random components. These very peculiar features of the data require modeling strategies that are not adequately handled using commonly employed methods, including panel data methods, and suggest that modeling patent data deserves further investigation.

Following Hausman *et al.* (1984) and Hall *et al.* (1986), the well-known HGH data, that is, the patents-R&D panel data of U.S. firms for the 1970-1979 period have been analyzed extensively in many studies such as Montalvo (1997), Blundell *et al.* (2002) and Guo and Trivedi (2002). The relative magnitudes of the estimated contemporaneous and lag effects vary somewhat across these studies depending on methodology. However, the main conclusion continues to be the one originally found by the first two studies. This being that, once you properly control for permanent differences in the propensity to patent across firms, there is very little direct evidence of anything but simultaneity in the year-to-year movements of patents and R&D.

[2] Recent studies include Blundell *et al.* (1995), Cincera (1997), Crepon and Duguet (1997), Montalvo (1997), Wang, Cokburn and Peterman (1998), Blundell *et al.* (2002), and Guo and Trivedi (2002).

Our dependent variable is the number of patents applied for by a particular firm during a given year that were eventually granted. As compared to the HGH data, the new sample for the U.S. manufacturing sector that we construct and analyze in this paper represents a longer data set over time, which allows us to explore dynamic effects by including more lags of number of patents. In addition, the new patent data are highly overdispersed, with much longer upper tails. These features clearly make the application of alternative estimation techniques even more desirable.

In implementation of the new semiparametric approach that we propose, the paper applies the methodology using the Jacobi polynomial series expansion. The variance parameter associated with unobserved heterogeneity is allowed to depend on covariates. The proposed series estimator provides flexible specifications for the conditional means, variances and covariances. In the application to patent activity, we also estimate various multiplicative individual effects models with predetermined regressors, including dynamic models, that have been developed recently by Chamberlain (1992), Wooldridge (1997), and Blundell *et al.* (2002).

Our empirical analyses show that, although results are somewhat sensitive to different estimation methods, the contemporaneous relationship between patenting and R&D continues being significant and rather strong, accounting for above 60% of the total R&D elasticity. For most of the distributed lag specification, the R&D elasticity of patents varies from 0.4 to 0.7, suggesting decreasing returns to scale. But unlike with the HGH data, the first (or the second) R&D lag appears to be as well significant; the associated coefficient has a value that is above 50% of the contemporaneous patents-R&D elasticity. In addition, the elasticity of current year's patenting with respect to R&D history is estimated to be around 0.17, irrespective of the lag length. These lag effects are higher than those previously found. The results might suggest that gestation lags in knowledge production have increased from the 1970's to the 1980's.

The remainder of this paper is organized as follows. Section 2 provides background on patents-R&D relationship, and describes our new data set. Section 3 presents various models particularly useful for the analysis of longitudinal patent data. In particular, this section develops a semiparametric generalization of a negative binomial-beta regression model. Section 4 discusses the empirical specifications and results. Section 5 concludes.

2 Background and data

Hausman *et al.* (1984) and Hall *et al.* (1986) investigated, among other things, whether there is a lag in the patent and R&D relationship. The former study analyzed patenting activity for 128 U.S. firms during the 1968-1975 period using up to 5 R&D lags. When they conditioned their estimates on the total number of patents received during the whole period, no coefficient except for the contemporaneous R&D variable were statistically significant either in Poisson or negative binomial count models. Hall *et al.* (1986) extended the sample in the cross-section and time-series dimensions. In particular, they considered 642 firms with patent and R&D data from 1972 to 1979. They also studied a subsample of 346 firms with a slightly larger time span covering 1970 to 1979. Using the same count-model frameworks as Hausman *et al.* (1984), the conclusion was again that, once you properly control for permanent differences in the propensity to patent across firms, there was very little direct evidence of anything but simultaneity in the year-to-year movements of patents and R&D.[3]

Recent studies have employed new estimation methods to try to deal with especial features of either the patent data or the patents-R&D relationship. Montalvo (1997), for example, addressed possible simultaneity problems in the Patent-R&D relationship. In particular, the consistency of the previously employed count models relied on the strict exogeneity of the expenditure in R&D with respect to patents. However, once a patent is granted, the firm may need to invest in R&D to transform the patent into a more commercial innovation in order to obtain benefits. From this viewpoint, R&D can be seen as a predetermined variable rather than strictly exogenous. Montalvo (1997) then proposed GMM estimation.[4] The main change in the results was that with GMM the HGH data delivered a significant first R&D-lag but an insignificant contemporaneous effect between patents and R&D. Thus results were inconclusive, and most likely a consequence of the high correlation among the R&D regressors. Blundell *et al.* (2002), in turn, produced quasi-differenced GMM estimators that allow for dynamic feedback from the history of the count process itself into the current patenting outcome. They found a contemporaneous relationship that was as strong as in

[3] Earlier work by Pakes and Griliches (1984) had already analyzed the data considered by Hausman *et al.* (1984) to try to identify the lag structure of the patent and R&D relationship. They did find lag effects, but with standard distributed-lags fixed-effect models that did not take the discreteness and non-negative nature of the patent data into account.

[4] This issue is also addressed by Hall *et al.* (1986) but in a different manner. They recognize that patents could be seen as an input to the R&D process rather than an output. To test this hypothesis, they perform a simple version of the Granger causality test, and conclude that "there may be simultaneous movements in patents and R&D, but there is little evidence that past success in patenting leads to an increase in a firm's future R&D." We have performed a similar test for our sample, with two to four lags of log R&D used to predict the current level of log R&D, including contemporaneous and lagged log patents (up to four lags) in the regression to see if they help to predict R&D in the presence of its past history. The result is that we neither find a clear effect from past patenting success into current R&D, although in our sample patents as an input to R&D can not be completely discarded. In particular, the first patent lag is the only lag that has an impact on R&D and goes in the right direction; but this is the case only when contemporaneous patents are left out of the regression. Contemporaneous patent activity, on the other hand, always helps to significantly explain current R&D. These results are available from the authors upon request.

previous work. In the context of cross section data, Guo and Trivedi (2002) estimated the patents-R&D relationship using flexible techniques that could better accommodate special features of the patent counts, especially their heavy upper tails and overdispersion. Results were again in line with Hall *et al.*'s (1986).

In this paper, we analyze a new data set using the proposed series estimator as well existing methods. Next, we describe the new sample and its construction process. As will become clear, the construction method follows the one used for the construction of the HGH data, which is described in Bound *et al.* (1984).

The new firm-level panel data set for the U.S. industrial sector covers the 1982-1992 period. The universe of the sample is the set of corporations and industry groups in the U.S. and Canadian manufacturing sector which existed in 1997 on Standard and Poor's Compustat Annual Industrial Files. From this sample frame, the subsample of 3034 U.S. firms that show strictly positive R&D expenses at least in one year is obtained for further subsampling. The Compustat files also provide each firm's book value of capital for each year, and the firm's standard industrial classification (SIC) and CUSIP identifies, where CUSIP (Committee on Uniform Security Identification Procedure) is the Compustat's identifying number for the firm. Patent numbers come from the U.S. Patent and Trademark Office. For the years 1971 to 1995, we obtain time series of utility patents granted to 8527 firms as distributed by year of application filing.

The matching of the Patent Office file and the Compustat data is no easy task. The difficulty is that the patenting organizations, although frequently corporations in our sample, may also be subsidiaries or have slightly different names from those given on the Compustat files. To do the matching, we proceed as follows. Out of the 3034 firms included in the R&D subsample, we first matched the organizations in the Patent Office file that have the same name (or slightly different name) as the ones in the Compustat data. In addition, for the firms in the Compustat file, we looked for their subsidiaries, and repeat the matching procedure.

The selection criterion for our sample is based on the absence of jumps, which coincides with the selection criterion for the Hausman *et al.*'s (1984) and Hall *et al.*'s (1986) data sets. More precisely, the final sample is chosen from the above universe by requiring that data on R&D investment, book value of capital, and patent counts are available for all years. We ended up with a balanced panel data sample composed of 391 U.S. firms with 11 years of data, 1982 to 1992, giving 4301 observations. Table 1 shows the distributions of net sales for firms in the universe and our sample. As in Hall *et al.* (1986), the organizations remaining in the sample show a size distribution heavily tilted toward the larger firms in the original universe. For example, out of the 2188 companies for which data on net sales in 1992 are available, our coverage of the largest firms is 68.0 percent, whereas it is 0.5 percent of the smallest. Regarding R&D expenditures, Table 2 shows that most of the firms excluded from the final sample were either smaller or did not report R&D during the 1982-1992 period. In fact, R&D spending is always strictly positive in our sample. The coverage of the larger R&D corporations is almost complete, and our sample includes 82.1 percent of the R&D dollars expended by the U.S. manufacturing sector between 1982 and 1992.

Table 3 reports summary statistics and descriptions for the variables included in the regressions. Figure 1 displays the frequency distribution of the number of patents applied for by firms. These show some striking features for the distribution of the number of patents. While the mean number of patents is relatively high (about 40 annual patents per firm during 1986-1992 period), the modal value of the number of patents is zero. The proportion of zero patents is about 16% during the study period, with proportion of firms with zero patents ranging from 12% in 1987 to 22% in 1983. The frequency distribution also shows that, on average, about 15% of the firms who patented during 1982-1992 did so only once per period, with the corresponding yearly figures ranging from a minimum of 11% in 1989 to a maximum of 19% in 1982. About 2.3% of the firms did not patent at all during the 1986-92 period. The distribution of the annual number of patents is highly right-skewed with range 0 to 1303 and median value of 5 patents. The third quartile and the 90th-percentile values are approximately 25 and 100 patents, respectively. The variance of the number of patents is quite large, which is consistent with the highly overdispersed nature of patent data used in recent studies. In comparison, for the 642-firm sample of the HGH data, the mean number of patents is 26.3 with standard deviation of 67.8, median of 4, and range of 0 to 906 patents. This means that our sample is more right-skewed and possesses a slightly larger overdispersion. These unique features of the data require special care in modeling patents, especially in the tails of the distribution.

3 Unobserved effects count data models

We examine various count panel data methods that are particularly useful for investigation of the relationship between the patenting process and R&D. In particular, we develop a series estimation approach that generalizes the familiar Negative binomial random effects model.

Consider a count panel data model with the conditional mean function

$$E(y_{it} \mid x_{it}, \nu_i) = \theta_{it}\nu_i, \quad i = 1, \dots, N; \quad t = 1, \dots, T, \tag{1}$$

where y_{it} is the observed value of the dependent variable for individual i at time t; $\theta_{it} = \exp(x'_{it}\beta)$; x_{it} is a $(p \times 1)$ vector of observed explanatory variables; β is the corresponding vector of parameters to be estimated; and ν_i is an unobserved individual effect. To develop further notation, let $y_{i.} = \sum_{t=1}^{T} y_{it}$ and $\theta_{i.} = \sum_{t=1}^{T} \theta_{it}$ denote sums over time, and $y_i = (y_{i1}, \dots, y_{iT})$.We shall focus on the case where N is large but T may not be.

3.1 Standard models

For multiplicative panel data model, the strict exogeneity assumption, conditional on ν_i, is given by

$$E(y_{it} \mid x_{i1}, \dots, x_{iT}, \nu_i) = \theta_{it}\nu_i \tag{2}$$

When x_{it} are strictly exogenous, the conditional maximum likelihood approach can be used to estimate β consistently. The conditional maximum likelihood approach, based on conditioning on $\sum_{t=1}^{T} y_{it}$ - which is the sufficient statistic for ν_i, allows for dependence between x_i and ν_i. Using the conditional maximum likelihood approach, Hausman *et al.* (1984) have proposed the Poisson and negative binomial fixed effects estimators.

In the application section, we also use the Poisson and negative binomial random effects models (Hausman *et al.* (1984). Here we focus on a general mixture model based on the negative binomial (Negbin) specification. The Negbin distribution with parameters (θ_{it}, δ), where again $\theta_{it} = \exp(x'_{it}\beta)$, is:

$$f(y_{it}) = \frac{\Gamma(y_{it} + \theta_{it})}{\Gamma(\theta_{it})\Gamma(y_{it} + 1)} \left(\frac{\delta}{1+\delta}\right)^{\theta_{it}} \left(\frac{1}{1+\delta}\right)^{y_{it}} \tag{3}$$

with $\text{mean}(y_{it}) = \frac{\theta_{it}}{\delta}$ and $\text{var}(y_{it}) = \frac{\theta_{it}}{\delta}\left(1+\frac{1}{\delta}\right)$. We specify:

$$\delta_i = \frac{\phi_i}{\nu_i},$$

where both ϕ_i and ν_i vary across individuals. Because of the presence of two error components, we adopt the following assumptions and parameterization. First, assume that ϕ_i and ν_i are independently distributed of each other so that δ_i is randomly distributed across individuals, independent of the x_{it}'s. Then, consider a new random variable

$$\begin{aligned}\xi_i &= \tfrac{\delta_i}{1+\delta_i} \\ &= \tfrac{\phi_i}{\phi_i+\nu_i}\end{aligned}$$

with density function $g(\xi_i)$, $0 < \xi_i < 1$. Then, assuming independence between y_{it}, y_{is} , conditional on $x_i = (x_{i1}, ..., x_{iT})$ and ξ_i, the joint density of y_i, given x_i, takes the following general form:

$$\begin{aligned}f(y_i \mid x_i) &= \int_0^1 \left[\Pi_{t=1}^T \tfrac{\Gamma(y_{it}+\theta_{it})}{\Gamma(\theta_{it})\Gamma(y_{it}+1)} {\xi_i}^{\theta_{it}} (1-\xi_i)^{y_{it}}\right] g(\xi_i)\, \mathrm{d}\xi_i \\ &= \left[\Pi_{t=1}^T \tfrac{\Gamma(y_{it}+\theta_{it})}{\Gamma(\theta_{it})\Gamma(y_{it}+1)}\right] \Xi(\theta_{i.}, y_{i.}),\end{aligned} \tag{4}$$

where

$$\Xi(\theta_{i.}, y_{i.}) = \left[E_\xi \left({\xi_i}^{\theta_{i.}} (1-\xi_i)^{y_{i.}}\right)\right]$$

Here $E_\xi[.]$ denotes expectation taken with respect to the distribution of ξ. It can be shown that, if ξ_i follows a beta distribution, then (4) reduces to the familiar Negbin random effects model, more precisely the Negbin-Beta mixture model, proposed by Hausman *et al.* (1984).

The Negbin-Beta model is based on arbitrary specifications of the density of the unobservable components. Section 3.3 presents series estimator of the random effects models given in (4) that does not require knowledge of the distribution of the unobservables.[5] Moment-based estimation approaches for mixture models are also available. In the context of random effects generalized linear models, Liang and Zeger (1986) and Zeger, Liang and Albert (1988) have proposed population-averaged mixed models in which serial correlations are allowed for but the random effects are averaged out. In the empirical section, using generalized estimating equations(GEE) approach, we estimate population-averaged panel data model based on the Negbin family.

3.2 Dynamic models

The models presented above have implicitly assumed strict exogeneity. We consider recently proposed methods that relax the strict exogeneity assumption (2). This includes methods that are applicable to estimate dynamic panel data models using generalized method of moments (GMM) framework.

Instead of (2), assume that

$$E(y_{it} \mid x_{i1}, ..., x_{it}, \nu_i) = \theta_{it}\nu_i, \quad t = 1, ..., T. \tag{5}$$

Chamberlain (1992) and Wooldridge (1997) have proposed GMM estimators for multiplicative panel data models, including count panel data models, that do not impose strict exogeneity assumption. In particular, they provide transformations that eliminate the fixed effects from model (5) by quasi-differencing and orthogonality conditions that can be employed for consistent estimation. The approach is applicable to distributed lag models with possible feedback and to models with lagged dependent variables. In the context of our application,

[5] In principle, a semiparametric approach based on Poisson-gamma baseline using Laguerre series expansion can be considered. In this paper, we shall focus only on series expansion approach based on Negbin-Beta baseline density for the counts.

the transformation that provides the appropriate residual functions is

$$r_{it} = y_{it} - \frac{\theta_{it}}{\theta_{it+1}} y_{it+1}, \quad t = 1, ..., T-1. \tag{6}$$

Let z_{it} be a q_t-vector of functions of $x_{i1}, ..., x_{it}$, $t = 1, ..., T-1$. Since the moment conditions $E(r_{it} \mid z_{it}) = 0$ hold, z_{it} is uncorrelated with r_{it}. This provides the basis for GMM estimation.

In the context of linear feedback model (LFM), Blundell, Griffith, and Windmeijer (2002) use variants of Chamberlain-Wooldridge moments conditions to estimate dynamic multiplicative individual effects models for count data. The mean function for dynamic model includes lagged dependent variable, which enters linearly, other conditioning variables in the exponential function, and the individual effects. For the case of one lag of the dependent variable, the conditional mean function for LFM is

$$E(y_{it} \mid x_{it}, \nu_i) = \beta_1 y_{it-1} + \exp(x_{it}^{*} \beta_2) \nu_i, \tag{7}$$

where x_{it}^{*} is a vector of other conditioning variables such that $x_{it} = (y_{it-1}\, x_{it}^{*})$.

In the empirical analysis, we estimate two versions of panel count data models using GMM framework. In the first specification, we use multiplicative distributed lag model, where contemporaneous and lags of regressors enter the exponential mean regression function. We also estimate a dynamic model similar to (7), where further lags of the dependent variable are included. Further details about choice of instruments and regressors will be given in the empirical section.

3.3 Semiparametric estimation

We now generalize the commonly used Negbin-beta model of Hausman *et al.* (1984). We develop semiparametric estimation methods for panel data models given in (4) that do not require knowledge of the distributions of ν_i. As in the Negbin-Beta mixture, the proposed model maintains the assumptions that the unobserved heterogeneity is independent of x_i, and that y_{it} and y_{is} are independent, conditional on x_i and unobserved heterogeneity. Following the techniques of Gallant and Nychka (1987) and Gurmu *et al.* (1999), the distribution of unobserved individual heterogeneity is estimated using series expansion. In the context of cross sectional count data models, Gurmu *et al.* (1999) use the generalized Laguerre series expansion around a gamma baseline density to model unobserved heterogeneity in Poisson mixture model. In this paper, we employ the Jacobi series expansion around a beta baseline density to approximate the distribution of unobserved individual effects in Negbin Mixture Model. Among other things, the ensuing Negbin-Jacobi mixture panel data model is more flexible with respect to the conditional mean and conditional variance/covariance.

Consider the mixture model (4). We approximate the density of $g(\xi_i)$ using series expansion around a beta distribution with parameters (a, b). The proposed approximate density is

$$g_N(\xi_i) = \frac{1}{B(a,b)\sum_{j=0}^{K} d_j^2}\xi_i^{a-1}(1-\xi_i)^{b-1}\left[\sum_{j=0}^{K} d_j h_{2j}^{-1/2} J_j(\xi_i)\right]^2 \tag{8}$$

where d_j's are constant coefficients in the polynomial expansion,

$$h_{2j} = \frac{j!\Gamma(a+j)\Gamma(a+b-1+j)\Gamma(b+j)}{(a+b-1+2j)\left(\Gamma(a+b-1+2j)\right)^2},$$

and

$$J_j(\xi_i) = \frac{\Gamma(a+j)}{(a+b-1+2j)} \sum_{l=0}^{j} \binom{j}{l} \frac{\Gamma(a+b-1+2j-l)}{\Gamma(b+j-l)\Gamma(j+1)} (\xi_i)^{j-l}$$

is the so called Jacobi polynomial of order j.[6] Similar to the Laguerre polynomials previously used in count and duration data literature, the Jacobi polynomials are orthogonal, and each with unit variance so that $\mathrm{var}(h_{2j}^{-1/2} J_j(\xi_i)) = 1$. The polynomials are squared to ensure that the density, $g_N(\xi_i)$, is positive. Since ξ_i takes values on the unit interval, the Jacobi polynomials seem to be the appropriate choice.

The next strategy is to determine $\Xi(\theta_{i.}, y_{i.})$ based on the approximate density $g_N(\xi_i)$.[7] After some algebra, we obtain

$$\Xi(\theta_{i.}, y_{i.}) = \frac{1}{\sum_{j=0}^{K} d_j^2} \sum_{j=0}^{K} \sum_{k=0}^{K} \sum_{l=0}^{j} \sum_{m=0}^{k} d_j d_k \Delta_{jk} \Psi_{lm} \ \mathrm{beta}(\theta_{.i} + a + j + k - l - m,\ y_{.i} + b), \quad (9)$$

where

$$\Delta_{jk} = (h_{j2} h_{k2})^{-1/2} \frac{\Gamma(a+j)\Gamma(a+k)}{\Gamma(a+b-1+2j)\Gamma(a+b-1+2k)}$$

and

$$\Psi_{lm} = (-1)^{l+m} \binom{j}{l} \binom{k}{m} \frac{\Gamma(a+b-1+2j-l)\Gamma(a+b-1+2k-m)}{\Gamma(a+j-l)\Gamma(a+k-m)}.$$

Inserting (9) into (4) gives the semiparametric (SPJ) density. For normalization, $d_0 = 1$. Thus, the log-likelihood function for the SPJ model is:

$$\mathcal{L}(\varphi_2) \quad = \textstyle\sum_{i=1}^{N} \sum_{t=1}^{T} \left[\log \Gamma(\theta_{it} + y_{it}) - \log \Gamma(y_{it} + 1) - \log \Gamma(\theta_{it}) + \frac{1}{T} \log(\Xi(\theta_{i.}, y_{i.})) \right], \quad (10)$$

where $\varphi_2 = (\beta' a\, b\ d_1\ \ldots\ d_K)'$ is the unknown parameter vector and $\Xi(\theta_{i.}, y_{i.})$ is given in (9). Since we have employed squared series expansion around beta distribution, the SPJ approach

[6] See Abramowitz and Stegun (1972) for definition of Jacobi Polynomials.

[7] Note that $\Xi(\theta_{i.}, y_{i.}) = \int_0^1 \xi_i^{\theta_{i.}} (1-\xi_i)^{y_{i.}} g_N(\xi)\, \mathrm{d}\xi_i$.

nests the Negbin-Beta mixture model of Hausman *et al.* (1984). Thus, if $d_1, ..., d_K = 0$ in (10), we obtain the log-likelihood function for the Negbin-beta panel data regression model.

In the application section, we use the Akaike information Criterion (AIC) to choose K. We also allow the variance parameter b associated with the unobservables to depend on covariates as described in the application section. As compared to the other models, the SPJ approach provides the most flexible specifications for the conditional means, variances, and covariances.[8] The approach is particularly useful in cases where explanatory variables satisfy the strict exogeneity assumption.

4 Empirical specifications and results

Using various individual effects count data models presented in the preceding section, we explore the relationship between R&D investment and patents at the firm level. The dependent variable that we take as an indicator of firms' technological output is the number of patents applied for by a particular firm in a given year that were eventually granted. The main explanatory variables of interest are the logarithms of current and past values of R&D expenditures in millions of 1983 dollars. Table 4 shows that lagged R&D expenditures are highly correlated over time.[9] The correlation between patent innovation and current or lagged R&D investments (in logs) is moderately high, on average about 0.56. As a measure of firm size, we use the logarithm of book value of capital in 1983 in millions of dollars as

[8] An appendix showing the first two moments of the standard panel count data and SPJ models is available from the authors. Consistency results for series estimators are considered by Gallant and Nychka (1987) and Gurmu *et al.* (1999)

[9] An additional evidence of the high correlation is given by the autoregressive structure of the log R&D series. We have performed AR regressions and found that a random walk process can not be rejected. Neither can an AR(2) process nor an AR(4) process, although estimated coefficients on the second, third and forth lags are relatively small. Other studies have also noted such high correlation and the associated computational problems; for instance, see Hall *et al.* (1986) and Cincera (1997).

a time-constant regressor. Another time-constant explanatory variable is a sector dummy, which equals 1 for firms in the scientific sector. To control for year effects, estimated models include year dummies as appropriate.

Starting from the general formulation in (1), the conditional mean in a distributed lag model is specified as

$$\begin{aligned} E(patent_{it} \quad | \quad & \log R\&D_{it}, ..., \log R\&D_{it-\tau}, w_{i,}, w_t^*, \nu_i) = \\ & \exp\left(\log R\&D_{it}\beta_1 + ... + \log R\&D_{it-\tau}\beta_{\tau+1} + w_i\gamma_1 + w_t^*\gamma_2\right)\nu_i, \end{aligned} \tag{11}$$

where w_i is a vector of firm specific effects such as book value of capital and w_t^* is a vector of time-specific variables, year dummies. As long as the expected number of patents, conditional on observables, is a scalar-multiple of the exponential mean form, the coefficient on $\log R\&D_{it-\tau}$ is an elasticity of the expected patent innovation with respect to R&D investment. The specification in (11) provides the framework for estimation of distributed lag model using fixed effects or random effects formulations outlined in section 3.

The GMM estimation method described in section 3.2. employ quasi-differencing conditions to eliminate fixed effect problems. Consequently, parameters on time-constant factors, such as the sectoral dummy and book value of capital in w_i, are not identified in LFM and related models. However, as noted by Wooldridge (1997), parameters on the interaction terms between time-varying and time-invariant regressors can be estimated. In the implementation of the GMM approach on distributed lag model, the regressors in x_{it} include $\log R\&D_{it}, ..., \log R\&D_{it-\tau}$ as well as the year dummies, say w_t, that are constant across firms. The GMM estimator applied to the distributed lag model uses the instruments

$$z_{it} = \left(1, \log R\&D_{it-(\tau+1)}, ..., \log R\&D_{i1}, w_t^*\right).$$

Alternatively, following Cincera (1997) and Crépon and Duguet (1997), we include additional instruments which result in restricted serial correlation. This involves adding past values of the dependent variable to the set of instruments z_{it}, resulting in

$$z_{it}^1 = \left(1, \log R\&D_{it-(\tau+1)}, ..., \log R\&D_{i1}, patent_{it-(\tau+2)}, ..., patent_{i1}, w_t^*\right)$$

The mean function underlying the estimation of the dynamic model emanating from (7) is

$$patent_{it-1}\omega_1 + ... + patent_{it-\kappa}\omega_\kappa + \exp\left(\log R\&D_{it}\beta_1 + w_t^*\gamma_2\right). \tag{12}$$

The ensuing GMM estimator[10] uses the instruments

$$z_{it}^2 = \left(1, patent_{it-(\kappa+1)}, ..., patent_{i1}, \log R\&D_{it-1}, ..., \log R\&D_{i1}, w_t^*\right).$$

The GEE approach adopts the exponential mean regression form (11) with assumed serial correlation structure. The results reported below are based on autoregressive correlation model of order 1, AR(1). The dispersion and correlation parameters in the weighting matrix are estimated iteratively using Pearson residuals. Finally, the semiparametric approach proposed in subsection 3.3 is implemented based on specification (11), and provides estimates of the all unknown parameters $(\beta_1, ..., \beta_{\tau+1}, \gamma_1, \gamma_2)$, along the estimates of the dispersion parameter and parameters in the series expansion. The dispersion parameter in the semiparametric log-likelihood function (10) is specified as $b = \exp(z_i'\alpha)$, where z_i is a vector of regressors consisting of over-time means of current and lagged values of log*R&D.*

Next, we estimate the empirical models that have been just described. The main results from the 8 models are given in Table 5. Although we have estimated patent equations with

[10] The Gauss code for the GMM estimator for the LFM model is obtained from Windmeijer (2002).

varying lags of log R&D and number of patents, only the preferred results with 3 or 4 lags are reported. If the coefficients on log R&D beyond lag 3 are insignificant at the 10% level, we simply report results from distributed lag specification of order 3. Comparisons of models are facilitated using the log-likelihood value, AIC, test statistics for overidentifying restrictions and serial correlation, and sum of the log R&D coefficients as appropriate.

The first four columns of Table 5 give results from standard Poisson and Negbin individual effects models. As compared to Poisson estimates, model comparison based on AIC favors the Negbin versions as expected. The Poisson-based estimate of the elasticity of patenting with respect R&D expenditure is about 0.65. In contrast, the elasticity estimates from the conditional Negbin and Negbin-Beta are much smaller, 0.40 and 0.46, respectively, and less precise, something expected given that the Negbin specification allows for an additional source of variance. These estimates are, in general, slightly larger than the ones based on 1970's data. For example, focusing on the fixed-effect models, which Hausman *et al.* (1984) find statistically preferred to the random-effects ones, these last authors find an elasticity of 0.43 and 0.38 for the conditional Poisson and conditional Negbin models, respectively. Hall *et al.* (1986) estimate the conditional Negbin model, and obtain 0.38 and 0.33 depending on the sample.

Our results also differ from theirs regarding the lag structure. Hausman *et al.* (1984) find a U-shaped lag structure with significant positive coefficients for t and $t-5$ for Poisson and Negbin random (uncorrelated) effects models, but only a contemporaneous relationship in their conditional fixed-effects version. Hall *et al.* (1986) also find only a significant contemporaneous relationship between patents and R&D with the conditional Negbin model.

Quite the contrary, we obtain U-shaped lag structures in both random- and fixed-effects Poisson-based models with significant positive coefficients for t, $t-1$ and $t-4$, and no evidence of it in the Negbin-based estimates. More important, in all models, the coefficients on contemporaneous as well as first-lag log R&D are positive and highly significant. In addition, the contribution of lagged R&D to current patenting activity is larger in our data set. In particular, focusing again in fixed-effect models, the conditional Poisson and conditional Negbin provide lagged-R&D contributions of 0.12 and -0.03 in Hausman *et al.* (1986), respectively; whereas these numbers in our case become 0.24 and 0.14. Both numbers are also bigger than the 0.05 found by Hall *et al.* (1986) for the conditional Negbin.[11]

Selected results from the semiparametric and moment-based estimators are given in columns 5 through 8 of Table 5. In terms of AIC, the semiparametric model dominates the likelihood-based mixture models, including the Negbin-Beta model which is nested in SPJ. The estimated R & D elasticities from the SPJ model are smaller than those obtained from the Negbin-Beta model. These random effects settings, including the SPJ framework, rely on the assumption that R&D expenditure variables in the patent equation are strictly exogenous. The GMM estimators considered in the last two columns of Table 5 relax the strict exogeneity condition. All GMM results are based on two-step estimation. For GMM results, m_1 and m_2 are the Arellano and Bond (1991) tests for first and second order serial correlation.[12] The GMM chi-square statistics for overidentifying restrictions and the tests for serial correlation all show that there is no clear evidence of misspecification.

[11] Montalvo (1997) estimates the conditional Poisson model for Hall *et al.*'s (1986) 346-firm sample. He finds a U-shaped lag structure with significant positive coefficients for t and $t-3$, and a lagged-R&D contribution of -0.01.

[12] The tests are asymptotically normally distributed. See Windmeijer (2002) for discussion of how the tests apply to the Chamberlain and Wooldridge residuals.

The results from the three distributed lag models show that elasticity varies from 0.37 for SPJ model to 0.67 for the GEE model, with quasi-differenced GMM estimate of elasticity lying in-between. For the LFM, the estimated elasticity is about 0.54, ignoring feedback. Consistent with previous studies, the contemporaneous partial effects of R&D on patenting are strong in all cases. The lag effects of R&D are smaller and, focusing on SPJ and GMM I results, the estimated overall lag effect is on the order of 0.13 to 0.17. Significantly positive effects of R&D occur at lag one for SPJ and at lag two for GMM I, but in the latter case negative and significant impact is found at lag four.

Given the last somewhat awkward result from GMM I, we feel it necessary to carry out a more detailed exploration with this technique, which we present in an appendix available on the Web.[13] The results show that, while the trade-off between contemporaneous and lag effects vary, the elasticity of current year's patenting with regard to R&D history is always estimated to be about 0.17, irrespective of the lag length. These results provide evidence that the impacts of R&D on patenting occur at an early stage of the R&D sequence. When we include additional lags the contribution of past R&D flips its sign and becomes negative. The analysis shows that, with 3 lags of R&D, significantly positive coefficients for contemporaneous and two-lags R&D, an elasticity with respect to R&D of 0.55, and a contribution of past R&D of 0.18. The total effect is similar to the one obtained with the HGH data, but the lag effect is larger. In particular, Montalvo (1997) applies GMM to the HGH data and obtains an elasticity of 0.56, and a lag effect of 0.15.

Results from the dynamic specification reported in the last column of Table 5 and in

[13] Detailed results from GMM I are available in a Web Table B2 at

an appendix (not given) show that the feedback effects of past patents on current patents are positive and significant at higher lags, but the results are sensitive to the number of lags included in the model. For the specification with four lags of the patent variable, the estimated overall effect is 0.33. When only one-period lag of patents is included, the coefficient on lagged dependent variable is insignificant, whereas the coefficient on log R&D is positive and highly significant.[14] In our specification of the LFM with $\kappa = 1$, the implied long run elasticity of patents with respect to R&D is about 0.59. Our analysis presented in appendix also shows that, although the magnitudes decline as more lagged values of the patent variable are added, the effects of log R&D on current year's patent generation remains positive and highly significant, even when we control for higher order lagged values of the dependent variable. We conclude that the results for LFM are largely consistent with the estimates obtained from GMM using distributed lag specifications.

5 Conclusions

This paper has investigated the impact of research and development and patent history on current patent activity using a firm level panel data set for the U.S. manufacturing sector from 1982 to 1992. The paper has also proposed a series estimator for count panel data models that generalizes the well-known Negbin-Beta mixture model. In addition, to address different unique features of patent and R&D data, we have estimated various distributed lag and dynamic count panel data models.

[14] In contrast, using HGH data and LFM esitmated by quasi-differencing approach with just one-period lag of patent included, Blundell *et al.* (2002) found puzzling results that the coefficient on lagged patent variable is negative, while the coefficient on log R&D is positive but insignificant. The authors attribute these results to a weak instruments problem due to persistence in both patents and R&D series. Our analysis, which uses data over longer horizon with the implied larger instrument set, shows that results from the dynamic feedback model are plausible.

The empirical analyses show that, although results are somewhat sensitive to different estimation methods, the contemporaneous relationship between patenting and R&D expenditures continues to be rather strong, accounting for over 60% of the total R&D elasticity. This conclusion is largely consistent with findings of previous studies. For most of the specifications, the overall R&D elasticity of patents varies from 0.4 to 0.7 suggesting decreasing returns to scale. This estimated elasticity is in most cases similar to the one obtained using the well-known 1970's HGH data. Our results differ mainly from the ones obtained with the HGH data regarding the lag structure of the patents-R&D relationship. In particular, unlike with the HGH data, we find the first or the second R&D lag, along with the contemporaneous effect, significant in all distributed lag specifications using flexible methods, with an associated coefficient whose value is approximately one-half of the contemporaneous patents-R&D elasticity. Moreover, while the trade-off between contemporaneous and lag effects vary, the elasticity of current year's patenting with respect to R&D history is estimated to be about 0.17, irrespective of the lag length. The lag effects are, therefore, moderately higher than have previously been found. Finally, results from linear feedback model are largely consistent with the estimates obtained from distributed lag specifications, and show that feedback effects of past patents on current patenting activity are positive and significant. All these results provide evidence that the impacts of R&D on patenting occur at an early stage of the R&D sequence.

In sum, comparing patenting activity in the U.S. manufacturing sector during the 1970s and the 1980s, we find that the overall long-run effect of R&D investment has not decreased over time, and that R&D history played a more important role during the 1980s than during the 1970s. The results could be interpreted as giving some additional support to studies, such as Hall (1993), that point out to a lower return to industrial R&D during the 1980s, because we find that gestation lags in knowledge production may have increased. On the other hand, if we take into account the 1980's R&D tax credit that generated incentives for firms to classify business costs as R&D expenditures (see also Hall, 1993), it is surprising that the average number of patents obtained for each dollar of R&D-classified investment has not declined, and our results are consistent with a more productive R&D activity. In fact, a larger contribution of R&D lags could certainly be due to a more time consuming R&D process, as well as a stronger dependence of current patenting on past successful R&D investment. Discriminating among these different interpretations is, we believe, an interesting and challenging task that we leave for future research.

References

Abramowitz M, Stegun IA (1972) Handbook of mathematical functions with formulas, graphs, and mathematical tables. U.S. Government Printing Office,Washington DC.

Acs ZJ, Audretsch DB (1989) Patents as a measure of innovative activity. Kyklos 2: 171-180.

Arellano M, Bond S (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Review of Economic Studies 58: 277-297.

Blundell R, Griffith R, Reenen JV (1995) Dynamic count data models of technological innovation. Economic Journal. 105: 333-344.

Blundell R, Griffith R, Windmeijer F (2002) Individual effects and dynamics in count data models. Journal of Econometrics 108: 113-131.

Bound J, Cummins C, Griliches Z, Hall BH, Jaffe A (1984) Who does R&D and who patents? In: Griliches R (ed) R&D, patents and productivity. University of Chicago Press, Chicago, 21-54.

Cameron CA, Trivedi PK (1998) Regression analysis of count data. Cambridge University Press, Cambridge.

Chamberlain G (1992) Comment: Sequential moment restrictions in panel data. Journal of Business & Economic Statistics10: 20-26.

Cincera M (1997) Patents, R & D, and technological spillovers at the firm level: Some evidence from econometric count models for panel data. Journal of Applied Econometrics 12: 265-280.

Crepon B, Duguet E (1997) Estimating the innovation function from patent numbers: GMM on count panel data. Journal of Applied Econometrics 12: 243-264.

Gallant A R, Nychka W (1987) Semi-nonparametric maximum likelihood estimation. Econometrica 55: 363-390.

Griliches Z (1990) Patent statistics as economic indicators: A survey. Journal of Economic Literature 28: 1661-1707.

Guo JQ, Trivedi PK (2002) Flexible parametric models for long-tailed patent count distributions. Oxford Bulletin of Economics and Statistics 63: 63-82.

Gurmu S, Rilstone P, Stern S (1999) Semiparametric estimation of count regression models. Journal of Econometrics 88: 123-150.

Gurmu S, Trivedi PK (1994) Recent developments in event count models: A survey. Thomas Jefferson Center Discussion Paper #261, Department of Economics, University of Virginia.

Hall BH, Griliches Z, Hausman JA (1986) Patents and R&D: Is there a lag? International Economic Review 27: 265-283.

Hausman JA, Hall BH, Griliches Z (1984) Econometric models for count data with applications to the patents R and D relationship. Econometrica 52: 909-938.

Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73: 13-22.

Montalvo JG (1997) GMM estimation of count-panel-data models with fixed effects and predetermined instruments. Journal of Business & Economic Statistics 15: 82-89.

Pakes A, Griliches Z (1984) Patents and R and D at the firm level: A first look. In: Griliches (ed) R&D, patents and productivity. University of Chicago Press, Chicago, 55-72 .

Stephan P, Audretsch D, Hawkins R (2000) The knowledge production function: Lessons from biotechnology. International Journal of Technology Managements 19: 165-178.

Wang P, Cockburn IM, Puterman ML (1998) Analysis of patent data - A mixed Poisson regression model. Journal of Business & Economic Statistics 16: 27-41.

Windmeijer F (2002) A Gauss programme for non-linear GMM estimation of exponential models with endogenous regressors for cross section and panel data. Institute for Fiscal Studies Working Paper Series No. CWP14/02, London.

Wooldridge JM (1997) Multiplicative panel data models without the strict exogeneity assumption. Econometric Theory 13: 667-678.

Zeger SL, Liang K-Y, Albert PS (1988) Models for longitudinal data: A generalized estimating equation approach. Biometrics 44: 1049-1060.

Table 1: Frequency Distribution of Net Sales in 1992 (Current Dollars)

	2188 R&D-firm Cross Section		Sample of 391 Firms		Coverage
Net sales	Number	Percent	Number	Percent	(%)
less than $1M	209	9.6	1	0.2	0.5
$1M - 10M	482	22.0	11	2.8	2.3
$10M - 100M	733	33.5	71	18.2	9.7
$100M - 1B	499	22.8	148	37.9	29.7
$1B - 10B	215	9.8	126	32.2	58.6
more than $10B	50	2.3	34	8.7	68.0

Source: Standard and Poor's Compustat Annual Industrial Files.

Table 2: R & D Expenditures in 1983 dollars for 1982-1992

	Data set 3034 firms	Sample 391 firms	Coverage (%)
Less than $1M	146.9	0.7	0.5
$1M-10M	3437.6	173.2	5.0
$10M-100M	23108.2	5760.0	24.9
$100M-1B	70144.4	46898.5	66.9
$1B-10B	257896.7	219637.0	85.2
More than $10B	165532.9	154890.2	93.6
total	520266.7	427359.6	82.1

Note: R & D represents all costs incurred during the calendar year that relate to the development of new products or services; the R & D deflator was provided by the U.S. Bureau of Economic Analysis.

Table 3: Summary Statistics

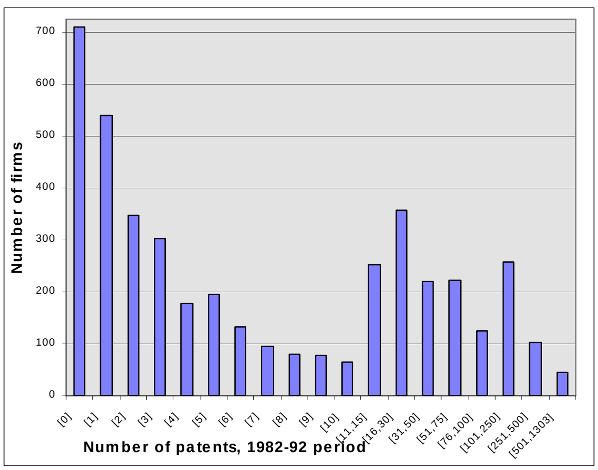
Variable	Mean	Standard Deviation	Minimum	Maximum
1982 - 1992:				
R&D Expenditure	98.9	332.8	0.010	4593
Number of patents	36.8	96.0	0	1303
First quartile number of patents	1			
Median number of patents	5			
Third quartile number of patents	24			
Proportion of zero patents	16.5			
Proportion with at least 100 patents	0.094			
Book value of capital in 1983	1,004.7	2895.5	0.115	29443
Fraction of firms in scientific sector	0.499	0.500	0	1
1986-1992:				
R&D Expenditure	109.4	368.0	0.053	4593
Number of patents	39.6	105.0	0	1303
First quartile number of patents	1			
Median number of patents	5			
Third quartile number of patents	26			
Proportion of zero patents	0.146			
Proportion with at least 100 patents	0.097			

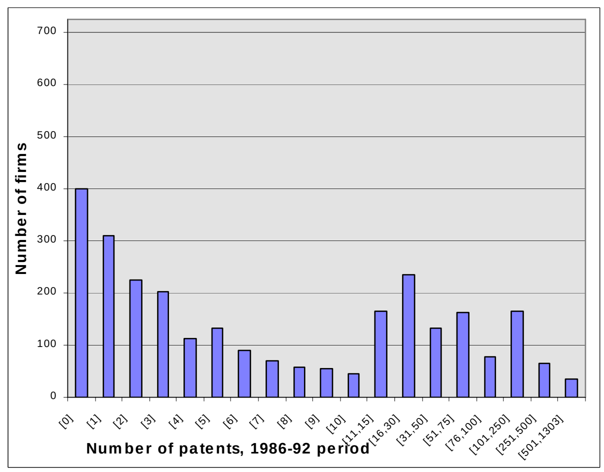
Note: R & D represents all costs incurred during the calendar year that relate to the development of new products or services in millions of 1983 dollars; Book value of capital (time-constant variable) is the firm's common equity liquidation value in millions of current dollars, and is based on calendar year end data ; Patents are utility patents granted to firms as distributed by year of application filing; The scientific sector (time-constant variable) is defined as firms in the drug, computer, scientific instrument, chemical and electric component industries). The R&D deflator was provided by the U.S. Bureau of Economic Analysis.

Table 4: Correlation Matrix

	Patent_{it}	log $\text{R\&D}_{it}$	log $\text{R\&D}_{it-1}$	log $\text{R\&D}_{it-2}$	log $\text{R\&D}_{it-3}$
log $\text{R\&D}_{it}$	0.561				
log $\text{R\&D}_{it-1}$	0.563	0.993			
log $\text{R\&D}_{it-2}$	0.563	0.983	0.992		
log $\text{R\&D}_{it-3}$	0.561	0.972	0.982	0.991	
log $\text{R\&D}_{it-4}$	0.560	0.960	0.971	0.981	0.992

Figure 1: Frequency Distribution of Patents





Note: The height of each bar gives number of firm-years corresponding to a given number of patents. The last 8 bars for the number of patents [11, 1303] are based on data of unequal intervals with interval widths of 5, 15, 15, 25, 25, 250, 250, and 500. Thus, the relative heights of the bars should be interpreted accordingly.

Table 5: Estimates of Knowledge Production Function from Various Models
Dependent Variable: Number of Patents Granted

Variable	(1) Conditional Poisson		(2) Conditional Negbin		(3) Poisson-Gamma		(4) Negbin-Beta	
log R&D$_t$	0.408	(14.96)[a]	0.257	(4.77)	0.408	(15.21)	0.289	(5.53)
log R&D$_{t-1}$	0.157	(4.48)	0.179	(2.46)	0.156	(4.45)	0.181	(2.52)
log R&D$_{t-2}$	-0.006	(0.17)	-0.022	(0.30)	-0.006	(0.18)	-0.010	(0.14)
log R&D$_{t-3}$	0.022	(0.64)	-0.016	(0.31)	0.021	(0.61)	0.001	(0.01)
log R&D$_{t-4}$	0.068	(2.67)			0.067	(2.64)		
Sum of R&D Elasticity	0.649		0.399		0.645		0.460	
Patent$_{t-1}$								
Patent$_{t-2}$								
Patent$_{t-3}$								
Patent$_{t-4}$								
Sum of Patent								
- Log-likelihood	8026.9		6198.6		10126.3		8379.0	
AIC	16075.8		12423.2		20282.5		16788.0	
GMM J-Statistics and [P-values]								

Table 5 (Continued)

Variable	(5) Semiparametric SPJ (K = 2)		(6) GEE with Serial Correlation		(7) GMM I^c		(8) GMM II^d	
log $\mathrm{R\&D}_t$	0.246	(4.62)	0.308	(4.12)	0.687	(4.00)	0.542	(13.19)
log $\mathrm{R\&D}_{t-1}$	0.176	(2.40)	0.220	(2.70)	-0.215	(0.98)		
log $\mathrm{R\&D}_{t-2}$	-0.028	(0.39)	0.066	(0.77)	0.322	(1.76)		
log $\mathrm{R\&D}_{t-3}$	-0.027	(0.50)	0.075	(0.97)	0.194	(1.01)		
log $\mathrm{R\&D}_{t-4}$					-0.478	(2.54)		
Sum of R&D Elasticity	0.369		0.669		0.511		0.542	
Patent_{t-1}							0.054	(0.84)
Patent_{t-2}							0.143	(2.17)
Patent_{t-3}							0.051	(0.84)
Patent_{t-4}							0.085	(1.98)
Sum of Patent							0.333	
- Log-likelihood	8306.4							
AIC	16646.8							
GMM J-Statistics and [P-values]					36.1 [0.278]		66.9 [0.281]	
m_1[P-values]					-3.63 [0.001]		-2.92 [0.004]	
m_2 [P-values]					0.08 [0.937]		-0.59 [0.555]	

[a] Absolute value of t-statistic.

[b] All models include year dummies and, except for fixed effects models, book value of capital and scientific sector dummy.

[c] Two-step quasi-differenced GMM estimator using z_{it}^1 as instruments.

[d] Two-step quasi-differenced GMM (LFM) estimator using z_{it}^2 as instruments.

Patents, R&D and Lag Effects: Evidence from Flexible Methods for Count Panel Data on Manufacturing Firms

Shiferaw Gurmu and Fidel Perez-Sebastian
Georgia State University and Univesidad de Alicante

March 2007

Appendices A and B
(Not for Publication)

Appendix A: Details on Unobserved Effects Count Data Models

Fixed Effects Methods

For the Poisson case, if $(y_{it} \mid x_{it}, \nu_i) \sim$ i.i.d. Poisson$(\theta_{it}\nu_i)$, then it can be shown that the conditional joint density for the *i-th* observation is

$$f(y_i \mid x_i, y_{i.}) = \frac{(y_{i}.)!}{\prod_{t=1}^{T} y_{it}!} \prod_{t=1}^{T} \left(\frac{\theta_{it}}{\theta_{i.}}\right)^{y_{it}}. \tag{A1}$$

The individual effects drop out upon conditioning. Hence, the conditional Poisson MLE of β can be obtained by maximizing the conditional likelihood function $\sum_{i=1}^{N} \log\ f(y_i \mid x_i, y_{i.})$, where $f(.)$ is given in (A1).[1] The first two moments of the conditional Poisson and other models presented below are given in Table B1 in Appendix B.[2]

Let ϕ denote the dispersion parameter of the negative binomial distribution such that $\delta_i^{-1} = \frac{\nu_i}{\phi_i}$. If $(y_{it} \mid x_{it}, \delta_i)$ is i.i.d. as negative binomial type 1 with mean θ_{it}/δ_i, the conditional maximum likelihood method can be used to estimate β. The Negbin conditional density for the *i-th* observation in which δ_i drops out can be obtained as

$$f(y_i \mid x_i, y_{i.}) = \frac{\Gamma(\theta_{i.})\,\Gamma(y_{i.}+1)}{\Gamma(\theta_{i.}+y_{i.})} \prod_{t=1}^{T} \frac{\Gamma(\theta_{it}+y_{it})}{\Gamma(\theta_{it})\,\Gamma(y_{it}+1)}. \tag{A2}$$

Given (A2), the ensuing log-likelihood function, which only involves β, is $\sum_{i=1}^{N} \log\ f(y_i \mid x_i, y_{i.})$. Table B1 shows that the mean of the conditional Negbin is the same as that of the conditional Poisson. As compared to Poisson fixed effects model, the Negbin fixed effects model has more general variance-covariance structure.

Mixture Models

Consider the Poisson specification with density

$$f(y_{it} \mid x_{it}, \nu_i) = \exp(-\theta_{it}\nu_i)\,(\theta_{it}\nu_i)^{y_{it}} / \Gamma(y_{it}+1),$$

where $\theta_{it} = \exp(x_{it}'\beta)$ as before. Assume that ν_i is independent of the observed covariates and that y_{it} and y_{is} are independent conditional on x_i and ν_i. Since ν_i is unob-

[1] The first order conditions for the conditional Poisson ML estimator is $\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \frac{y_{i.}}{\theta_{i.}}\theta_{it}\right)x_{it}$. See Cameron and Trivedi (1998) and references there in for a discussion on how these conditions have been used as a basis of estimation using the method of moments.

[2] Except for the conditional Poisson, the moments for the other models have not been provided explicitly in the literature.

servable, we need to integrate it out. This gives the Poisson random effects (Poisson-Gamma) model:

$$f(y_i \mid x_i) = \left[\Pi_{t=1}^{T} \frac{\theta_{it}^{y_{it}}}{\Gamma(y_{it}+1)}\right] \frac{\Gamma(y_{i.}+\alpha)}{\Gamma(\alpha)} \alpha^{-y_{i.}} \left(1+\frac{\theta_{i.}}{\alpha}\right)^{-(\alpha+y_{i.})}, \tag{A3}$$

The unknown parameter vector for Poisson-Gamma mixture model is $(\beta\ \alpha)$. In the Poisson random effect specification with $E(y_{it} \mid x_i, \nu_i) = \exp(x_{it}'\beta)\nu_i$, if x_{it}'s are constant, we have randomness only across individuals. If x_{it}'s are constant, there is no variation across time. This is a potential problem with the Poisson random effects model given in (A3). We consider alternative models that exhibit randomness both across individuals and across time.

The most common distribution for ξ_i in (4) of the paper is the beta density

$$g_b(\xi_i) = \frac{1}{B(a,b)} \xi_i^{a-1} (1-\xi_i)^{b-1}, \tag{A4}$$

where $B(.)$ is the beta function:

$$B(\gamma,\omega) = \frac{\Gamma(\gamma)\Gamma(\omega)}{\Gamma(\gamma+\omega)}.$$

In this case, it can readily be shown that

$$\Xi(\theta_{i.}, y_{i.}) = \frac{B(\theta_{i.}+a, y_{i.}+b)}{B(a,b)},$$

and (**??**) reduces to the Negbin random effects model:

$$f(y_i \mid x_i) = \left[\Pi_{t=1}^{T} \frac{\Gamma(y_{it}+\theta_{it})}{\Gamma(\theta_{it})\Gamma(y_{it}+1)}\right] \frac{B(\theta_{i.}+a, y_{i.}+b)}{B(a,b)} \tag{A5}$$

The unknown parameters of the Negbin-Beta model are (β, a, b). The moments given in Table B1 shows that the Negbin-Beta specification is more flexible than that of the Poisson-gamma specification.

In the context of random effects generalized linear models, Liang and Zeger (1986) and Zeger, Liang and Albert (1988) have proposed population-averaged mixed models in which serial correlations are allowed for but the random effects are averaged out. In the empirical section, we estimate population-averaged panel data model based on the Negbin family using GEE. The estimation approach is based on the first two moments of the Negbin distribution, incorporating unobserved effects and serial correlation. As before, let y_i (θ_i) denote a $T \times 1$ vector with *t-th* element y_{it} (θ_{it}). The first-order conditions take the form

$$\sum_{i=1}^{N} D' V_i \,(y_i - \theta_i) = 0, \tag{A6}$$

where $D = [diag(\theta_{it})x_i]$ is a $T \times p$ matrix and V_i is a $T \times T$ weighting matrix involving the mean parameters, correlation parameters , and dispersion or scale parameter ϕ. Depending on the assumed correlation structure, the dispersion parameter as well as the correlation parameters can be estimated iteratively using Pearson residuals.

Moments for SPJ Model

To develop the moments associated with the SPJ model, define

$$\varpi(\delta_1, \delta_2) = \frac{1}{\sum_{j=0}^{K} d_j^2} \sum_{j=0}^{K} \sum_{k=0}^{K} \sum_{l=0}^{j} \sum_{m=0}^{k} d_j d_k \Delta_{jk} \Psi_{lm} \ \mathrm{B}(\delta_1 + a + j + k - l - m, \ \delta_2 + b) \quad \text{(A7)}$$

for arbitrary constants $\delta_1$and δ_2. The first and second order moments of the SP2 density are also shown in Table B1.

Appendix B: Tables

Table B1: Moments of Some Panel Count Data Models

Model	Mean$(y_{it} \mid x)$ Or Mean$(y_{it} \mid x, \, y_{i.})$	Var$(y_{it} \mid x)$ Or Var$(y_{it} \mid x, \, y_{i.})$	Cov$(y_{it}, y_{is} \mid x)$ Or Cov$(y_{it}, y_{is} \mid x, \, y_{i.})$
Conditional Poisson	$\theta_{it} \frac{y_{i.}}{\theta_{i.}}$	$y_{i.} \left(\frac{\theta_{it}}{\theta_{i.}}\right) \left(1 - \frac{\theta_{it}}{\theta_{i.}}\right)$	$-\, y_{i.} \left(\frac{\theta_{it}}{\theta_{i.}}\right) \left(1 - \frac{\theta_{is}}{\theta_{i.}}\right)$
Conditional Negbin	$\theta_{it} \frac{y_{i.}}{\theta_{i.}}$	$y_{i.} \left(\frac{\theta_{it}}{\theta_{i.}}\right) \left(1 - \frac{\theta_{it}}{\theta_{i.}}\right) \left(\frac{y_{i.} + \theta_{i.}}{1 + \theta_{i.}}\right)$	$-\, y_{i.} \left(\frac{\theta_{it}}{\theta_{i.}}\right) \left(1 - \frac{\theta_{is}}{\theta_{i.}}\right) \times \left(\frac{y_{i.} + \theta_{i.}}{1 + \theta_{i.}}\right)$
Poisson-Gamma	θ_{it}	$\theta_{it} + \alpha^{-1}\theta_{it}^2$	$\alpha^{-1}\theta_{it}\theta_{is}$
Negbin-Beta	$(a-1)^{-1}\, b\theta_{it}$	$\frac{(a+b-1)(a+\theta_{it}-1)b\theta_{it}}{(a-1)^2(a-2)}$	$\frac{(a+b-1)b\theta_{it}\theta_{is}}{(a-1)^2(a-2)}$
SPJ	$\theta_{it}\varpi(-1,1)$	$\theta_{it}\left[\varpi(-1,1)\left(1-\theta_{it}\varpi(-1,1)\right) + \left(1+\theta_{it}\right)\varpi(-2,2)\right]$	$\theta_{it}\theta_{is}\left[\varpi(-2,2) - \varpi^2(-1,1)\right]$

Note: The moments for the conditional Poisson and Negbin are obtained by additionally conditioning on $y_{i.} = \sum_{t=1}^{T} y_{it}$.

Table B2: Estimates from GMM I^c Specifications (Details)

Variable	(1)		(2)		(3)		(4)		(5)	
log R&D$_t$	0.403	(5.24)	0.441	(4.83)	0.373	(2.87)	0.687	(4.00)	0.692	(2.68)
log R&D$_{t-1}$	0.176	(2.23)	-0.096	(0.76)	-0.063	(0.40)	-0.215	(0.98)	-0.313	(1.22)
log R&D$_{t-2}$			0.234	(2.58)	0.386	(2.42)	0.322	(1.76)	0.488	(1.96)
log R&D$_{t-3}$					-0.140	(1.05)	0.194	(1.01)	0.259	(1.27)
log R&D$_{t-4}$							-0.478	(2.54)	-0.236	(0.64)
log R&D$_{t-5}$									-0.398	(1.03)
Sum of R&D Elasticity	0.579		0.580		0.555		0.511		0.493	
GMM J-Statistics and [P-values]	88.2 [0.225]		65.14 [0.335]		53.6 [0.177]		36.1 [0.278]		22.3 [0.272]	

[a] Absolute value of t-statistic.

[b] All models include year dummies.

[c] Two-step quasi-differenced GMM estimator using z_{it}^1 as instruments.

Table B3: Estimates from GMM II[c] Specifications (Details)

Variable	(1)		(2)		(3)		(4)		(5)	
log R&D$_t$	0.599	(19.55)	0.563	(17.39)	0.534	(14.38)	0.542	(13.19)	0.485	(10.43)
Patent$_{t-1}$	0.036	(0.94)	0.039	(0.79)	0.069	(1.25)	0.054	(0.84)	0.192	(2.49)
Patent$_{t-2}$			0.114	(2.67)	0.177	(3.10)	0.143	(2.17)	0.091	(1.03)
Patent$_{t-3}$					0.157	((2.92)	0.051	(0.84)	0.054	(0.77)
Patent$_{t-4}$							0.085	(1.98)	0.153	(2.70)
Patent$_{t-5}$									-0.012	(0.22)
Sum of Patent	0.036		0.152		0.402		0.333		0.477	
GMM J-Statistics	105.3		89.1		76.4		66.9		49.3	
and [P-values]	[0.265]		[0.359]		[0.371]		[0.281]		[0.463]	

[a] Absolute value of t-statistic.

[b] All models include year dummies.

[c] Two-step quasi-differenced GMM (LFM) estimator using z_{it}^2 as instruments.

Table B5: Summary Statistics for Patent Numbers - Comparison with HGH Data

Statistic	1982-1992 sample	1972-1979 sample (HGH Data)
Mean	36.8	26.3
Standard deviation	96.0	67.8
First quartile	1	1
Median	5	4
Third quartile	24	19
Fraction of zeros	0.16	0.23
Fraction of at least 100	0.10	0.07

Description: Patents are utility patents granted to firms as distributed by year of application filing.

Table B6: Autoregressive Estimates for log R&D

Equation	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
log R&D$_{t-1}$	0.994 (433.560)	1.121 (60.451)	1.118 (59.828)	1.116 (60.007)	1.110 (59.572)	1.107 (59.357)	1.099 (58.891)	1.115 (60.000)	1.108 (59.632)
log R&D$_{t-2}$		−0.128 (6.885)	−0.107 (3.917)	−0.116 (4.258)	−0.129 (6.988)	−0.126 (6.760)	−0.114 (6.123)	−0.130 (6.999)	−0.120 (6.441)
log R&D$_{t-3}$			−0.018 (1.013)	0.088 (3.339)					
log R&D$_{t-4}$				−0.095 (5.443)					
log Patent$_t$					0.019 (3.155)	0.023 (3.680)	0.026 (4.099)		
log Patent$_{t-1}$					−0.001 (0.231)	0.005 (0.755)	0.013 (1.882)	0.016 (2.721)	0.025 (3.854)
log Patent$_{t-2}$						−0.013 (2.068)	0.002 (0.316)	−0.005 (0.835)	0.010 (1.399)
log Patent$_{t-3}$							−0.009 (1.254)		−0.006 (0.895)
log Patent$_{t-4}$							−0.022 (3.623)		−0.022 (3.505)
$\bar{R}^2$	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986	0.986

Notes:

1. Absolute values of t-statistic are in parenthesis.
2. All equations contain a separate intercept for each year.
3. We added 1/3 to the patent variable before taking the log due to the presence of zeros.